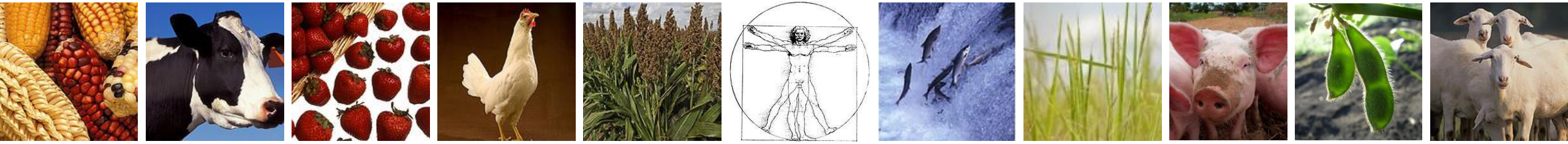


Submission of Genomes to GenBank

Karen Clark

NCBI Genome Resources Workshop

PAG XXVII January 14, 2019

U.S. National Library of Medicine
National Center for Biotechnology Information

GenBank is…

- regular data exchange
- data standards
- open and unrestricted access

- globally comprehensive coverage
- scientific database of record
- public forum for the scientific process

Why GenBank?

- Accessioned datasets for journals and so that everyone is referring to the same sequences
- Common point of data access, regardless of species
- Common file formats, regardless of species
- Value added: QA on incoming submissions, and the errors are reported back so that they can be corrected

What is needed?

- BioProject to describe the research effort
- BioSample to describe the sample that was sequenced
- Assembled sequences
- Assignment information, if relevant = which sequences belong to or *are* chromosomes
- AGP to assemble scaffolds into chromosomes, if relevant

WEB: https://www.ncbi.nlm.nih.gov/genbank/eukaryotic_submission/

Factsheet: https://go.usa.gov/xEbbF (https://ftp.ncbi.nlm.nih.gov/pub/factsheets/Factsheet_EukGenomeSubmission.pdf)

BioSample: Metadata is important

- Use the same BioSample for data that come from the same source
  - 1 BioSample for the DNA reads and the genome assembly created from those reads
  - RNAseq data would have a different BioSample for each tissue
- When reads from multiple BioSamples are used to create a genome assembly, make a new ‘combination’ BioSample that has the common information and refers to the others in a note

Pictures from Ocellaris_clownfish wikipedia

“Model organism or animal sample” package

https://submit.ncbi.nlm.nih.gov/biosample/template/?package=Model.organism.animal.1.0&action=definition

Required.

One is Required.

Optional.

*sample_name | *organism | *sex | *tissue

strain | isolate | breed | cultivar | ecotype

age | dev_stage

sample_title | bioproject_accession

biomaterial_provider | birth_date | birth_location | breeding_history | breeding_method | cell_line

cell_subtype | cell_type | collected_by | collection_date | culture_collection | death_date | disease

disease_stage | genotype | geo_loc_name | growth_protocol | health_state | isolation_source

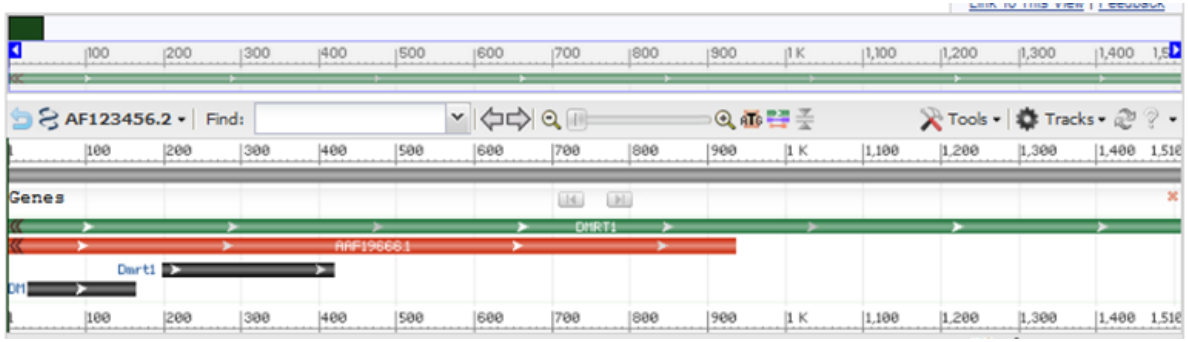
lat_lon | phenotype | sample_type | specimen_voucher | store_cond | stud_book_number | treatment | description

Plus ‘custom attributes’, if desired

Improvements to genome submissions

- We accept Gapped submissions:
  - FASTA sequence of the scaffolds, with information about which Ns represent gaps
  - Can be submitted with an AGP file to make chromosomes
- Still accept the traditional type:
  - contigs + optional AGP file(s) to make scaffolds and/or chromosomes
- Batch submission available:
  - A single submission can have up to 400 genomes that meet certain requirements, eg that they are in the same BioProject

Annotation

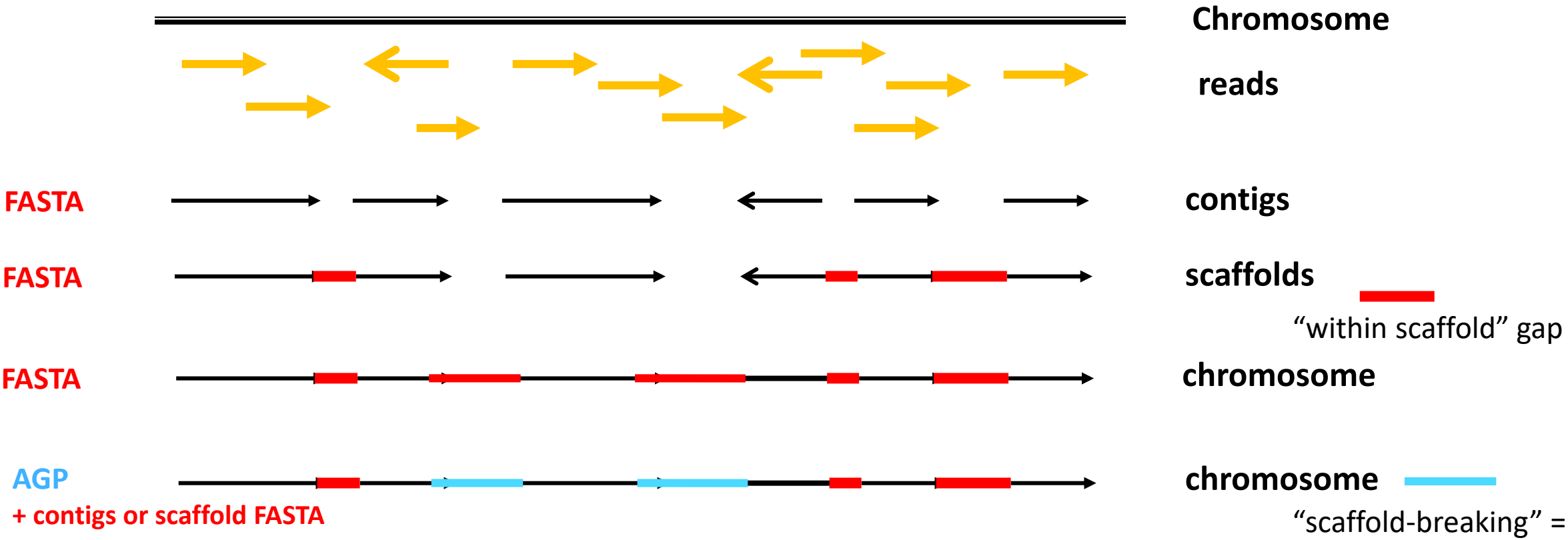


- Optional, but nice to have
- Should be on the upper-level objects, ie scaffolds or chromosomes
- Product names should conform to the International Protein Nomenclature Guidelines (NCBI/SwissProt/EBI)
  - https://www.ncbi.nlm.nih.gov/genome/doc/internatprot_nomenguide/
- Input can be
  - 5-column feature table (.tbl file)
    - https://www.ncbi.nlm.nih.gov/genbank/eukaryotic_genome_submission_annotation/
  - GenBank-specific GFF/GTF file
    - https://www.ncbi.nlm.nih.gov/genbank/genomes_gff/
- Run our tools to create the file for submission

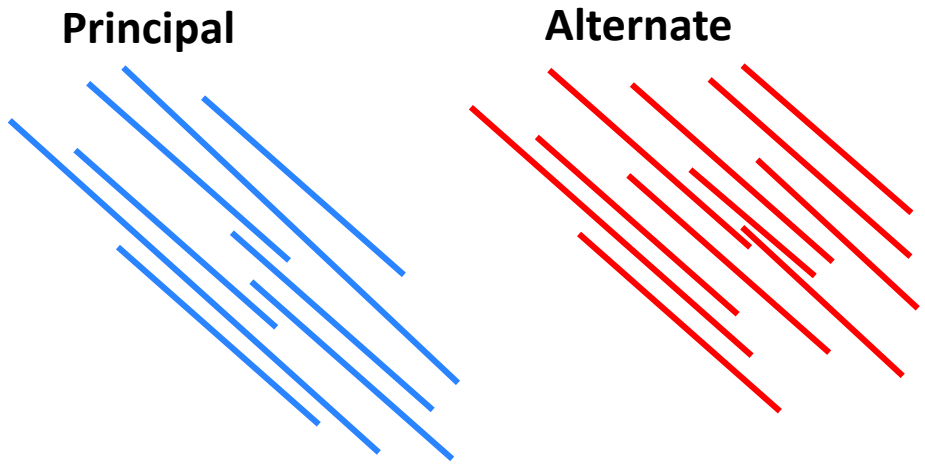
What to submit when unannotated?

FASTA if no scaffold-breaking gaps

+AGP file to assemble chromosomes that have scaffold-breaking gaps (AGP optional for ‘within scaffold’ gaps)



“Diploid” assemblies-



- Submit each pseudohaplotype to GenBank as a separate assembly
- Use the same BioSample for both
- Create a separate BioProject for each
- Indicate in the comment box whether this is
  - the principal assembly
  - the alternate pseudohaplotype assembly
- Encode distinguishing information in the assembly names, eg

Organism	Principal	Alternate
Bos taurus	Btau_diploid_p1.0	Btau_diploid_a1.0
T. guttata	bTaeGut1_v1.p	bTaeGut_v1.h

- An umbrella BioProject for the pair will be created
- The assemblies will be linked to each other in the Assembly resource
- We’re still working on the tools to submit and to view these more easily

Example BioProjects

VGP male Taeniopygia guttata (zebra finch) genome sequencing and assembly Accession: PRJNA510143 ID: 51014

Taeniopygia guttata (zebra finch) genome sequencing and assembly

Accession	PRJNA510143
Type	Umbrella project
Organism	Taeniopygia guttata [Taxonomy ID: 59729] Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Archelosauria; Archosauria; Dinosauria; Saurischia; Theropoda; Coelurosauria; Aves; Neognathae; Passeriformes; Passeroidea; Estrildidae; Estrildinae; Taeniopygia; Taeniopygia guttata
Grants	"Molecular mechanisms of vocal learning" (Grant ID 1, Howard Hughes Medical Institute)
Submission	Registration date: 14-Dec-2018 Vertebrate Genomes Project - G10K
Relevance	Model Organism

NAVIGATE UP

This project is a component of the Vertebrate Genomes Project

NAVIGATE ACROSS

3 additional projects are components of the Vertebrate Genomes Proje

Project Data:

Resource Name	Number of Links
SEQUENCE DATA	
Nucleotide (total)	35
WGS master	2
PUBLICATIONS	
PubMed	2
PMC	2
OTHER DATASETS	
BioSample	1
Assembly	2

This project encompasses the following 2 sub-projects:

Project Type	Number of Projects
Genome sequencing	
Highest level of assembly :	
Chromosomes	1
Scaffolds	1
Total	2

BioProject accession	Assembly level	Organism	Title
PRJNA489098	Chromosomes	Taeniopygia guttata	Taeniopygia guttata (zebra finch) genome s…eing and assembly, primary haplotype, v1 (Vertebrate Genomes Project)
PRJNA489099	Scaffolds	Taeniopygia guttata	Taeniopygia guttata (zebra finch) genome sequencing and assembly, alternate haplotype, v1 (Vertebrate Genomes Project)

Assembly | Advanced | Browse by organism

Summar… | Sort by Significance | Download Assemblies

L…ks from BioProject

…tems: 2

Filters activated: Exclude anomalous. Clear all

bTaeGut1_v1.p

1. Organism: Taeniopygia guttata (zebra finch)
Sex: male
Submitter: Vertebrate Genomes Project
Date: 2018/12/20
Assembly level: Chromosome
Genome representation: full
GenBank assembly accession: GCA_003957565.1 (latest)
RefSeq assembly accession: n/a
IDs: 2174221 [UID] 8109868 [GenBank]
BioProject BioSample Nucleotide INSDC Taxonomy WGS Master

bTaeGut1_v1.h

2. Organism: Taeniopygia guttata (zebra finch)
Sex: male
Submitter: Vertebrate Genomes Project
Date: 2018/12/20
Assembly type: alternate-haplotype
Assembly level: Scaffold
Genome representation: full
GenBank assembly accession: GCA_003957525.1 (latest)
RefSeq assembly accession: n/a

…ect/PRJNA510143

Accession: PRJNA489098 ID: 48909…

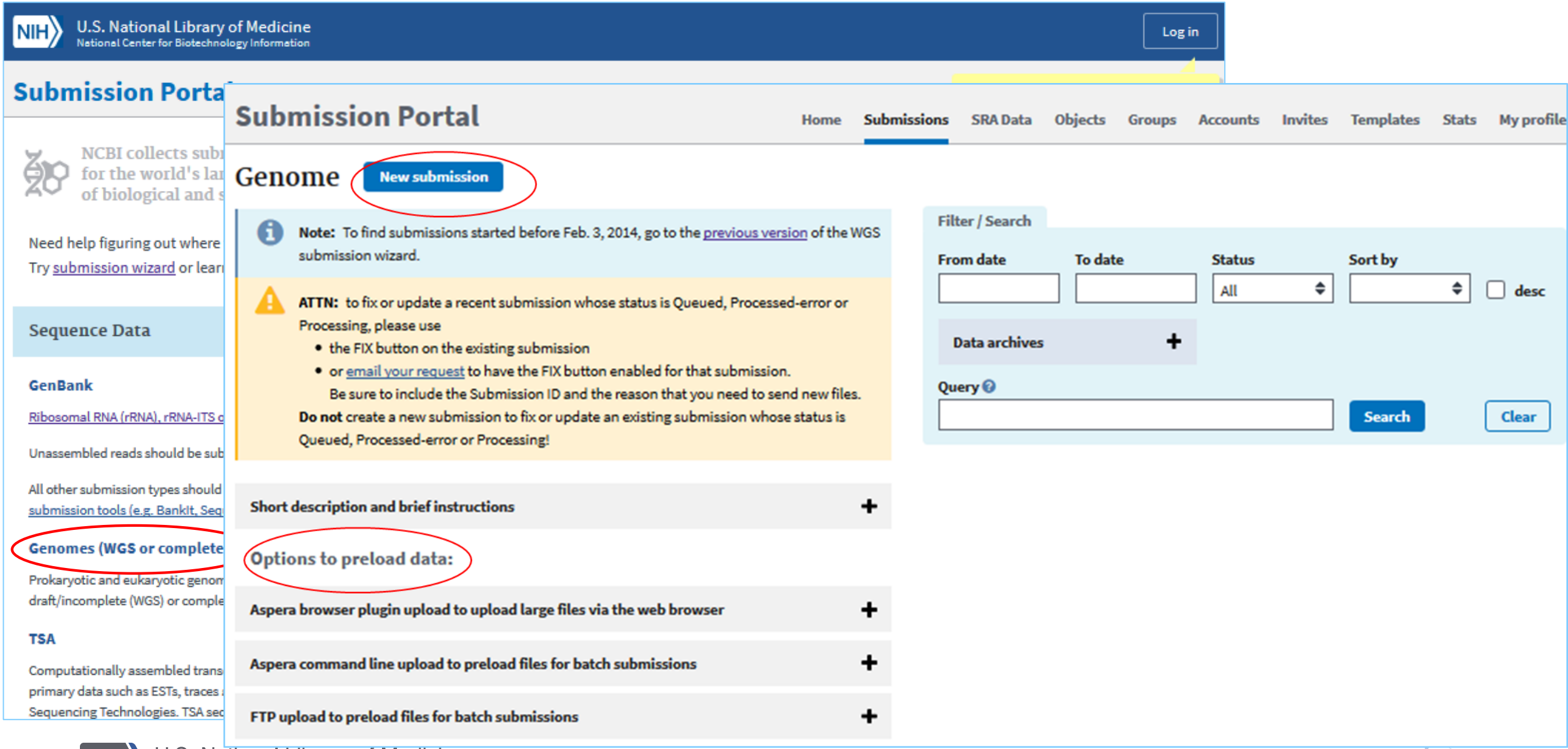
…aplotype, v1

OTHER DATASETS	
BioSample	1
Assembly	1

Assembly details: Download

Assembly	Level	WGS	Chrs	BioSample	Taxonomy
GCA_003957565.…	Chromosome	RRCB00000000	33	SAMN02981239	Taeniopygia guttata

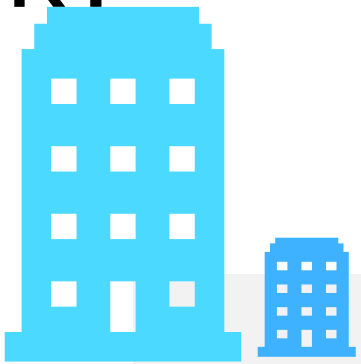
https://submit.ncbi.nlm.nih.gov



What happens after an assembled genome is submitted to NCBI/GenBank?

QA

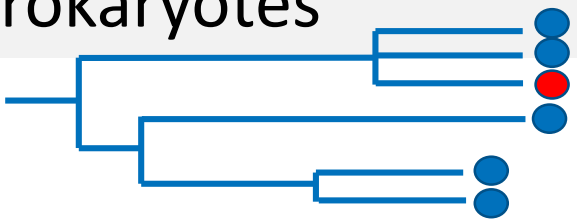
Foreign Contamination Screen



Genome size:
too big or too small?

Validation or
Discrepancy
Problems

Average Nucleotide Identity (ANI) analysis of prokaryotes



Sources of contamination

Sample Collection → DNA Extraction → Library Preparation → Sequencing → Submission to database

In vivo

Infection with viruses, microbes or parasites

Symbionts

Endosymbionts

Commensals

Food

Gut microbes

Cross-contamination of samples

Errors in sample tracking

Contaminated reagents

Aerosols

Mitochondrial DNA

Cross-contamination of DNA

In vitro

Adaptors & primers

Spike-ins

Cloning vectors

E. coli & phage

IS & Transposons

Human handlers

Errors in tracking of libraries

Errors in tracking of DNA preps

Spill-over

Carry-over

In silico

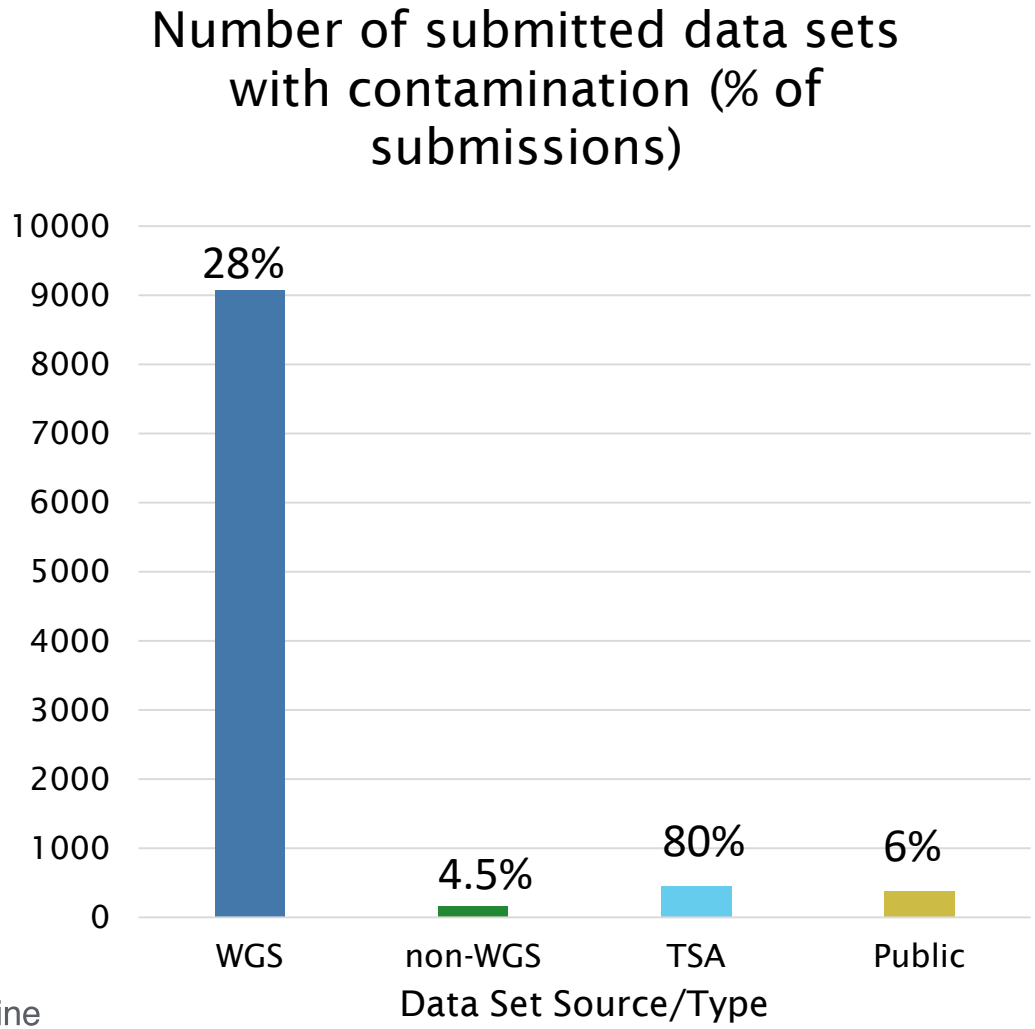
Errors in tracking data or metadata

Impact of contamination

- Functions & Homologies
  - Zebrafish gene that was actually from contaminating mouse sequence
- Propagation of Errors
  - Gene ended up in treefam.org as a zebrafish gene in the middle of mouse/rat genes, although it's not present in the current zebrafish asm
- Errors in Clinical Metagenomics
  - Identification of the reads & assemblies from a metagenome rely on a clean reference set
- Personally Identifiable Information
  - Human sequences that become public in an unrelated genome assembly

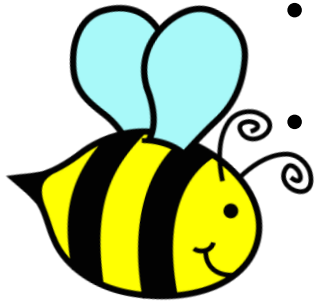
Image from the Worth1000
Chimaera 5 Photoshop contest

How often is contamination found?



January - November 2018

Project to screen and clean our public assemblies



I. Foreign Contamination Screen:

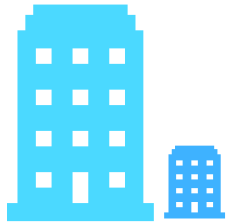
- Screen is updated as more genomes and more adaptors are added
- Suppress sequences that are all contamination
- Convert contaminating sequences to ‘contamination gaps’ to retain the coordinate system
- For example, removed138 contigs from Bombus impatiens (GCA_000188095.3) AND 5 contigs from its symbiont Candidatus Schmidhempelia bombi (GCA_000471645.2); cross-contaminants

II. ANI analysis to confirm the organism of prokaryotic genomes.

- For example, JVTR00000000 (GCA_001053395.1) was changed from 'Enterobacter cloacae' to 'Enterobacter kobei’

We notify the submitter of our findings before making any changes

Genome Size Test



- The genome size is generally expected to be within 4 standard deviations of the median size of the genomes of that species that are already in GenBank.
  - Failure: Contamination in vivo or in vitro; file mixup
- Genome submissions could pass this test just because there are not yet enough genomes of that species in GenBank.
- Example:

Genome size: 31,983,140
Average: 304,945,999. expected range: [152,473,000-457,418,998]
ERROR: size too small but based on only 2 samples

Updates?

- New sequencing and new assembly -> new version of WGS
  - New accessions for contigs and scaffolds; same accessions for chromosomes
- Reassemble existing sequences
  - New accessions for new scaffolds; same accessions for old scaffolds and for chromosomes
- Reannotate
  - Submitter should track the gene locus_tag’s and protein accessions to the new assembly

Thank you.

This work was supported by the Intramural Research Program of the NIH, National Library of Medicine.

GenBank
Shelby Bidwell
Larissa Brown
Jianli Dai
Scott Durkin
Michel Eschenbrenner
Linda Frisse
Leigh Riley
Karen Clark
Ilene Mizrachi

BioProject / Biosample
John Anderson
Carol Scott
Tanya Barrett

GEO
Emily Clough
Carlos Evangelista
Irene Kim
Pierre Ledoux
Hyeseung Lee
Kimberly Marshall
Katherine Phillippy
Patti Sherman
Stephen Wilhite
Tanya Barrett

GEO developers
Alexandra Soboleva
Maxim Tomashevsky
Nadezhda Serova
Naigong Zhang

Annotation Pipeline
Francoise Thibaud-Nissen
Paul Kitts
Mike Dicuccio
Wratko Hlavina
Avi Kimchi

Jinna Choi
Patrick Masterson
Eyal Mozes
Robert Smith
Alexandre Souvorov

RefSeq/Gene
Eric Cox
Catherine Farrell
Tamara Goldfarb
Diana Haddad
John Jackson
Vinita Joardar
Kelly McGarvey
Michael Murphy
Nuala O'Leary

RefSeq Developers
Alex Astashyn
Olga Ermolaeva
Vamsi Kodali
Craig Wallin

Shashi Pujar
Bhanu Rajput
Sanjida Rangwala
Lillian Riddick
Barbara Robbertse
Brian Smith-White
Pooja Strope
Anjana Vatsan
David Webb

GDV/Remap/GBench
Valerie Schneider
Peter Meric
Nathan Bouk
Hsiu-Chuan Chen
Cliff Clausen
Anatoliy Kuznetsov

A cast of thousands
Ken Katz
Michael Ovetsky
Lukas Wagner
Andrei Shkeda
Donna Maglott
Kim Pruitt
Jim Ostell

Watch NCBI News for updates!

http://www.ncbi.nlm.nih.gov/news/
https://www.youtube.com/user/NCBINLM

NCBI Genome Resources Workshop

Time	Topic
12:50 – 1:10	Submission of Genomes to GenBank *Karen Clark*
1:10 – 1:30	GEO Submissions and Usage *Steve Wilhite*
1:30 – 1:55	From Annotation to Visualization: Exploring Genes and Genomes with NCBI Tools *Eric Cox*
1:55 – 2:15	Programmatic Access to Genomic Data: E-Utilities and FTP *Vamsi K. Kodali*
2:15 – 2:35	NCBI Resources for Phyletically-Defined Next Generation Analysis in and out of the Cloud (a.k.a. Cool New Stuff!) *Ben Busby*
2:35 – 3:00	Q & A session